Propaganda Detection using NLP

Abstract

This paper explores propaganda detection using NLP, addressing two tasks: identifying propaganda in text and classifying its technique. We compare Logistic Regression and BERT models to evaluate effectiveness. Our approach aims to offer a practical solution for detecting deceptive content, supporting media transparency and informed public discourse

1. Introduction

The automatic detection of propaganda techniques in text is a vital task in natural language processing (NLP), particularly for mitigating the spread of misinformation across digital platforms. This project addresses two key subtasks: **Task 1 — Technique Classification**, where the model is given a known propagandistic span and must predict its technique; and **Task 2 — Span Detection and Technique Labelling**, where the system must extract propagandistic spans from a sentence and assign each one the correct technique.

For **Task 1**, we implemented two models: a baseline **TF-IDF + Logistic Regression classifier**, and a contextual **BERT-based classifier** fine-tuned on labelled spans and their surrounding sentences. For **Task 2**, we explored three complementary approaches. First, a **rule-based system** that uses keyword patterns and syntactic cues to match propaganda phrases and their techniques. Second, a **BERT token classification model** trained using BIO tagging to identify span boundaries and label them at the token level. Third, a **BERT-based question answering (QA) classifier**, which treats span detection as a QA task by querying for each technique and extracting corresponding spans from the sentence. This multi-approach strategy allows us to compare the interpretability, coverage, and effectiveness of rule-based heuristics and deep learning methods under a unified evaluation framework.

2. Related Work

Propaganda detection has evolved from document-level classification to fine-grained span-level identification. Early efforts (Rashkin et al., 2017) labelled entire news sources, but recent work (Da San Martino et al., 2019) introduced span-level annotations of specific propaganda techniques, improving interpretability and precision. Rule-based systems such as Proppy (Barrón-Cedeño et al., 2019) leveraged lexical features, while neural approaches adopted contextual embeddings. BERT-based models have shown strong performance in token classification and question answering, both relevant to span detection. Our work builds on these foundations by combining rule-based heuristics, BIO-tagged token classification with BERT, and QA-style span extraction to address propaganda detection at multiple levels.

3. Methodology

Propaganda is a form of strategic communication aimed at influencing public perception, attitudes, or behaviour by selectively presenting facts, emotional appeals, or persuasive techniques. It often seeks to promote a specific agenda or ideology, using rhetorical devices to manipulate opinion rather than encourage critical thinking.

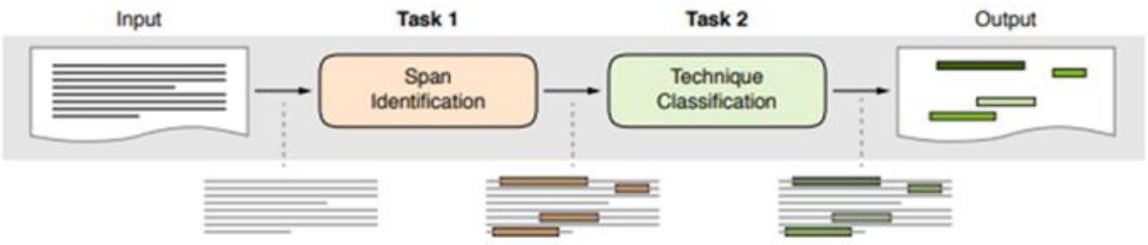


Figure 1: The full propaganda identification pipeline

3.1 Data Preparation: The dataset comprises sentences annotated with one of eight propaganda techniques or a "not propaganda" label. Propagandistic spans are marked using <BOS> and <EOS> tags, which were parsed to extract span-text and associated labels. For **Task 1**, spans were cleaned, lowercased, and used in two models: a TF-IDF + Logistic Regression classifier and a BERT-based sequence classifier. For **Task 2**, full sentences were tokenized, and BIO tags (B-technique, I-technique, O) were generated for each token using Hugging Face's tokenizer. The BERT token classifier and QA-style model used this format, while the rule-based method applied regex patterns after Stopword removal.

3.1.1 Exploratory Data Analysis (EDA): An initial analysis of the dataset was conducted to understand its composition. The label distribution showed class imbalance, with techniques like loaded language and name calling/labelling occurring most frequently, while causal simplification and repetition appeared less often. Span length distribution revealed that most spans were between 2 and 6 tokens long, although outliers existed. Word clouds and frequency plots highlighted the dominance of emotionally charged terms, validating the presence of rhetorical devices typically associated with propaganda. These observations guided decisions around sequence padding, vocabulary curation, and model input structuring.

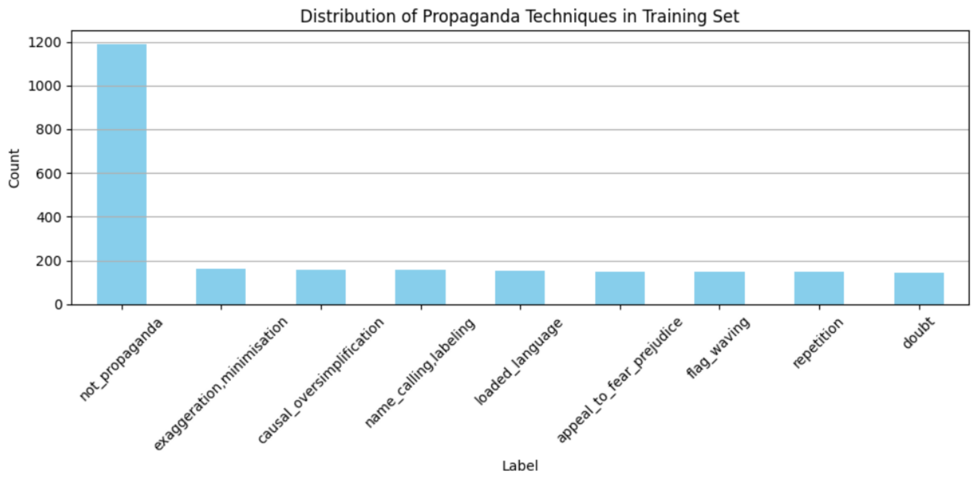


Figure 2 Propaganda Techniques in Training Set

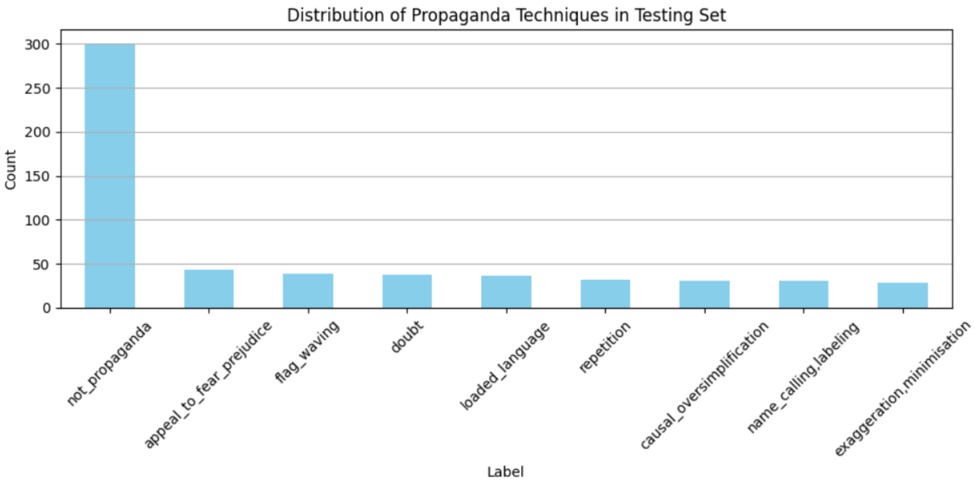
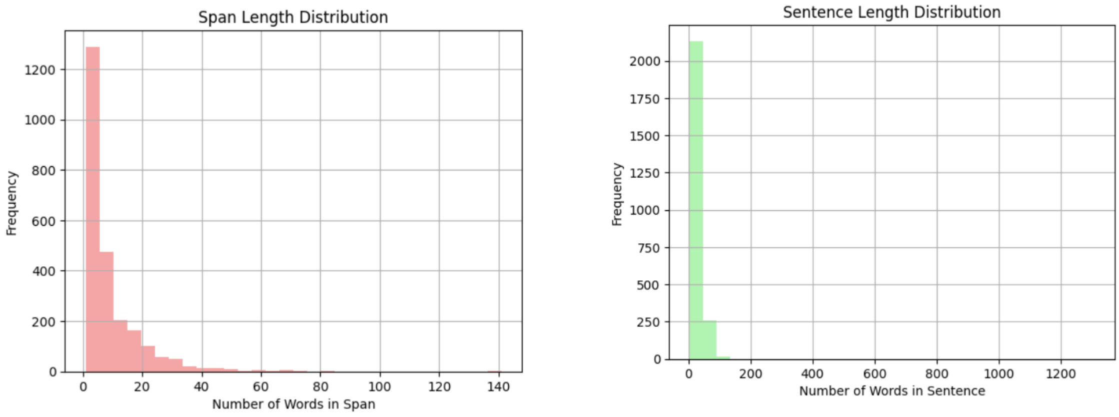


Figure 3 Propaganda Techniques in Testing Set

Span and Sentence Length Analysis



3.2 Task 1: Propaganda Technique Classification: In Task 1, the objective was to classify a given propagandistic span into one of eight predefined propaganda techniques. We implemented and evaluated two models for this task: a traditional machine learning baseline using TF-IDF features with logistic regression, and a contextual deep learning model using a fine-tuned BERT transformer.

3.2.1 TF-IDF + Logistic Regression: The baseline model applies a bag-of-words approach, using scikit-learns TfidfVectorizer to extract unigram and bigram features from the text. The input to the model consists of the propagandistic span concatenated with its surrounding sentence context, providing additional linguistic cues. The vocabulary size was limited to 5,000 terms to prevent overfitting and reduce sparsity, and all text was normalized to lowercase. A Logistic Regression classifier was trained using a regularization parameter C=1.0C = 1.0C=1.0 and class weights were set to 'balanced' to compensate for class imbalance among the eight propaganda techniques.

This model was evaluated on a held-out validation set and achieved an accuracy of 35.5% and a macro-averaged F1-score of 0.35. It performed reasonably well on some techniques such as flag waving (F1 = 0.56) and doubt (F1 = 0.50), but exhibited poor performance on others, particularly name calling and loaded language, with F1-scores below 0.20. These results suggest that the model was able to learn distinguishable patterns for a few dominant techniques, but struggled to generalize across semantically overlapping or less frequent categories. While this approach provides a simple and interpretable baseline, it lacks the capacity to model nuanced rhetorical signals or contextual semantics.

3.2.2 BERT Sequence Classification: To overcome the limitations of feature sparsity and limited context modelling, we implemented a deep learning approach using the Bert-base-uncased model from Hugging Face. For each instance, the input was constructed by concatenating the propaganda span and its surrounding context sentence, separated by the [SEP] token. Inputs were tokenized using BertTokenizer, and padded or truncated to a maximum length of 128 tokens. The model was fine-tuned using BertForSequenceClassification for 3 epochs, with a learning rate of 5e-5 and batch size of 16. The AdamW optimizer was used along with CrossEntropyLoss for multi-class classification.

This BERT-based model achieved an improved accuracy of 50.6% and a macro F1-score of 0.50, significantly outperforming the baseline. Notably, it achieved strong performance on name calling (F1 = 0.60), exaggeration/minimisation (F1 = 0.57), and flag waving (F1 = 0.59).

Despite these gains, the model still struggled with loaded language (F1 = 0.25), which remains a challenging category due to its abstract nature and reliance on subtle connotations

3.2.3 Summary: In summary, the BERT sequence classifier demonstrated substantial improvements over the TF-IDF + Logistic Regression baseline across most techniques. Its ability to leverage contextual embeddings allows it to distinguish between techniques with greater nuance and accuracy. However, both models showed limitations on less frequent or semantically diffuse categories, highlighting the need for further data balancing, advanced modelling (e.g., attention mechanisms), or ensemble methods to boost performance.

3.3 Task 2: Span Detection and Technique Labelling: To evaluate the effectiveness of span detection and technique labelling in Task 2, we assessed each model using precision, recall, F1-score, and accuracy. We also reported span-level metrics such as Span Match Accuracy, Technique Accuracy, and Full Match Accuracy to capture how well each system identified both the boundaries and the correct label of propaganda spans.

3.3.1 BERT Token Classification (BIO Tagging): The BIO-tagging model, implemented using BertForTokenClassification, achieved a **token-level accuracy of 63%** on the validation set. It performed strongly on the dominant O (non-propaganda) class, with **precision of 0.69**, **recall of 0.89**, and **F1-score of 0.78**, reflecting its bias toward non-propaganda tokens due to class imbalance. Among propaganda techniques, **appeal to fear/prejudice** achieved the highest F1-score (0.31), followed by **causal oversimplification** (0.22) and **doubt** (0.24). In contrast, techniques such as **loaded language**, **name calling/labelling**, **not propaganda**, and **repetition** received F1-scores of 0.00, indicating the model's inability to detect them. This failure is likely due to their lower frequency and greater linguistic subtlety, which makes learning effective representations difficult. The **macro-averaged F1-score was 0.19**, while the **weighted average F1-score was 0.56**, largely influenced by the high number of correct predictions for the O class. These results reveal a core limitation of the BIO tagging approach: it tends to Favor majority classes and struggles with nuanced or rare rhetorical patterns. Future work should explore class rebalancing, span-aware loss functions, or hybrid models to address this imbalance and improve label-level generalization.

3.3.2 BERT QA-Based Classifier: The QA-based model reframed propaganda span detection as a series of technique-specific question-answering tasks. For each technique, the model received a fixed query (e.g., "What phrase reflects flag waving?") and returned a span prediction or no answer if the technique was not found. This approach supports multi-label and multi-span detection within the same sentence. Evaluation showed that the model achieved a **Span Match Accuracy of 44.14%**, indicating it correctly identified span boundaries in nearly half the cases. However, **Technique Accuracy** dropped to **6.55%**, and **Full Match Accuracy**—requiring both correct span and label—was only **4.83%**. This reveals that while the model can often locate relevant phrases, it struggles to assign the correct propaganda label. The QA model performed well for long or explicit spans but suffered from low recall, particularly with subtle or overlapping techniques. Its design—treating each technique in isolation—may limit contextual understanding when multiple techniques co-occur. Despite these limitations, the model's precision-oriented output makes it suitable for settings where minimizing false positives is more important than exhaustive coverage. Future work may focus on improving prompts, span merging, and recall-oriented strategies to enhance overall utility.

3.3.3 Rule-Based Classifier: The rule-based classifier used manually defined patterns and keyword lists for each propaganda technique. While interpretable and computationally efficient, it performed poorly with **precision of 0.006**, **recall of 0.0034**, **F1-score of 0.0044**, and **accuracy of 0.22%**. It failed to handle varied phrasing, subtle cues, and overlapping techniques, often producing false positives from overly broad patterns. Although useful as a baseline, the model's inability to generalize highlights the limitations of static pattern matching and underscores the necessity for context-aware, data-driven approaches in complex linguistic tasks.

3.3.4 Comparative Summary: Among the three approaches, the **BIO tagging model** provided the best balance of span localization and classification but struggled with minority classes. The **QA-based model** offered better precision on clear spans but suffered from low recall and mislabelling. The **rule-based classifier**, while fast and interpretable, performed the worst due to its rigid pattern matching. Overall, the results highlight the challenges of multi-span, multi-label propaganda detection and the need for models that combine contextual understanding with span-aware prediction strategies.

4. Hyperparameter

In this study, hyperparameter settings were intentionally fixed across all models to ensure a consistent and fair comparison between approaches. For Task 1, the TF-IDF + Logistic Regression model used a fixed vocabulary size of 5,000 and included both unigrams and bigrams to capture short-term context. The regularization parameter was set to C = 1.0, and class weights were balanced to address label imbalance across propaganda techniques. These settings reflect standard practice in text classification and were selected to serve as a strong and interpretable baseline.

The BERT sequence classification model was fine-tuned using the Bert-base-uncased configuration for three epochs, with a batch size of 16 and a learning rate of 5e-5. These values were chosen based on prior benchmarks for transformer models and were not tuned further to maintain fairness.

For Task 2, the rule-based model required no trainable parameters; instead, fixed keyword lists and regular expressions governed its logic. The BIO-tagging model also used standard fine-tuning settings identical to Task 1, with padding tokens excluded from loss computation. The QA-based model followed the same training setup, using technique-specific queries without prompt engineering. These consistent hyperparameter choices prioritized interpretability and reproducibility across all experiments.

5. Result and Evaluation Metrics

In this study, we adopt a comprehensive evaluation framework to assess the performance of our models in both Task 1 and Task 2. The primary metrics used include **precision**, **recall**, and **F1-score**—standard in classification tasks, especially when dealing with imbalanced datasets or multi-label outputs such as those found in propaganda detection.

- **Precision**

  Precision measures the proportion of predicted propaganda instances that are actually correct. It reflects the model's ability to avoid false positives. A high precision score indicates that when the model predicts a propaganda span or technique, it is likely to be correct.

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

- **Recall**

  Recall evaluates the proportion of actual propaganda instances that the model successfully identifies. It measures the model's sensitivity to propaganda cues and reflects its capacity to minimize false negatives.

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

- **F1-Score**

  The F1-score is the harmonic mean of precision and recall. It provides a single balanced measure of a model's performance, particularly useful when precision and recall are in tension.

$$\text{F1 Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

These metrics were computed for each class (propaganda technique) as well as macro and weighted averages, ensuring a nuanced and fair evaluation across both frequent and rare labels.

5.1 Task 1: Propaganda Technique Classification: Two models were implemented for Task 1: a TF-IDF + Logistic Regression baseline and a fine-tuned BERT sequence classification model. The **TF-IDF + Logistic Regression** model achieved an overall **accuracy of 35.5%** and a **macro F1-score of 0.35**. It performed best on techniques like flag waving (F1 = 0.56) and doubt (F1 = 0.50), but struggled with less frequent or ambiguous labels. The **BERT classifier** outperformed the baseline with an improved **accuracy of 50.6%** and **macro F1-score of 0.50**. It showed notable gains in techniques like name calling (F1 = 0.60), exaggeration/minimisation (F1 = 0.57), and flag waving (F1 = 0.59). These results reflect the model's ability to leverage contextual information for nuanced label distinctions.

5.2 Task 2: Propaganda Span Detection and Labelling

Three approaches were evaluated for Task 2: a rule-based system, a BERT BIO tagging model, and a BERT QA-based model. The **Rule-Based Classifier** demonstrated extremely poor performance, with **precision = 0.006**, **recall = 0.0034**, and **F1-score = 0.0044**, reflecting its inability to generalize beyond fixed keyword patterns. While interpretable and computationally simple, it was ineffective for complex, varied input and served primarily as a baseline. The **BERT Token Classification (BIO Tagging)** model achieved a **token-level accuracy of 63%**, with a **macro F1-score of 0.19** and **weighted F1-score of 0.56**. It performed strongly on non-propaganda tokens (O class) but failed to detect several techniques, such as loaded language, repetition, and name calling, which received F1-scores of 0.00. The **QA-Based Classifier** achieved **similar span-level results**, with **44.14% span match accuracy**, **6.55% technique accuracy**, and **4.83% full match accuracy**. It was more precise on longer, clearly defined spans but showed lower recall and missed overlapping techniques due to its single-query-per-technique formulation.

6. Analysis

The evaluation across both tasks highlights the effectiveness of contextual models and the challenges posed by class imbalance and span-based prediction. In **Task 1**, the BERT sequence classification model outperformed the TF-IDF + Logistic Regression baseline, achieving an accuracy of 50.6% compared to 35.5%. BERT's contextual embeddings enabled better recognition of subtle propaganda techniques, whereas the TF-IDF model performed better on lexically distinct classes but lacked semantic understanding.

For **Task 2**, the span detection and technique labelling task presented substantial challenges. The **BIO tagging model** achieved a **token-level accuracy of 63%**, with a **macro F1-score of 0.19**. It performed well on the dominant O (non-propaganda) class but failed to detect several underrepresented techniques such as loaded language, name calling/labelling, repetition, and not propaganda, all of which had F1-scores of 0.00. This highlights the model's sensitivity to class imbalance The **QA-based classifier** achieving **44.14% span match accuracy**, **6.55% technique accuracy**, and **4.83% full match accuracy**. While it performed better for longer, explicit spans, it failed to capture overlapping or subtle techniques, likely due to its one-query-per-technique setup.

In contrast, the **rule-based model**, despite being interpretable and lightweight, performed poorly across all metrics, with an **F1-score of just 0.0044** and **accuracy of 0.22%**, confirming its ineffectiveness for complex linguistic patterns

7. Conclusion

This study explored multiple approaches for propaganda detection, tackling two tasks: classifying known propaganda spans (Task 1) and detecting both spans and techniques within full sentences (Task 2). For Task 1, we compared a traditional TF-IDF + Logistic Regression model with a fine-tuned BERT classifier. The BERT-based model significantly outperformed the baseline, demonstrating the effectiveness of contextual embeddings in identifying nuanced rhetorical techniques.

Task 2 posed greater challenges due to the need for accurate span localization and multi-label classification. Among the models evaluated, the BERT BIO tagging model achieved the most balanced performance, with strong token-level accuracy but limited success in technique labelling. The QA-based classifier struggled with recall and overlapping techniques. The rule-based system, although interpretable, proved ineffective due to its reliance on rigid patterns.

Overall, BERT-based models consistently outperformed traditional and rule-based methods, but all approaches faced difficulties with minority classes and subtle techniques. This suggests that future work should explore data augmentation, hierarchical modelling, and more robust handling of class imbalance. Improving label differentiation and incorporating external knowledge sources may also enhance model robustness in real-world applications of propaganda detection.

8. Future Work

While our experiments demonstrated that BERT-based models consistently outperform traditional approaches in both tasks, there is still considerable room for improvement. In **Task 1**, classifying known propaganda spans worked well overall, but misclassification of less frequent techniques remains a concern. Future work could explore advanced fine-tuning strategies, better sampling methods to balance the training data, or ensemble models that combine strengths from multiple classifiers.

For **Task 2**, the BIO tagging and QA-based approaches struggled to assign correct technique labels even when span boundaries were reasonably well identified. This suggests the need for **multi-stage models** that first detect candidate spans and then classify them using specialized sub-models. Alternatively, **joint learning frameworks** that simultaneously optimize span detection and classification may offer improved coordination between the two objectives.

Additionally, **prompt engineering** and **few-shot learning** could enhance the QA model's ability to generalize across techniques, especially when examples are limited. Exploring architectures like **Span BERT** or incorporating **external knowledge bases** (e.g., rhetorical structures or propaganda lexicons) could further improve performance.

Finally, **explainability** and **real-world adaptability** should be prioritized in future models to support deployment in sensitive contexts such as journalism, policy-making, and social media monitoring.

References

- Da San Martino, G., Yu, S., Barrón-Cedeño, A., Petrov, R., & Nakov, P. (2019). Fine-grained analysis of propaganda in news articles. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP), 5636–5646.
- Traylor, R., Caliskan, A., & Harwell, D. (2019). Bias in propaganda detection systems: A case for fairness and interpretability. arXiv preprint arXiv:1911.04474.
- Rashkin, H., Choi, E., Jang, J. Y., Volkova, S., & Choi, Y. (2017). Truth of varying shades: Analyzing language in fake news and political fact-checking. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2931–2937.
- Barrón-Cedeño, A., Da San Martino, G., & Nakov, P. (2019). Proppy: Organizing the news based on their propagandistic content. Information Processing & Management, 56(5), 1849–1864.
- Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of NAACL-HLT, 4171–4186.

- Torok, R. (2015). Developing an explanatory model for the process of online radicalisation and terrorism. Security Informatics, 4(1), 1–10.
- Miller, C. (1939). How to Detect Propaganda. The Institute for Propaganda Analysis.

Weston, A. (2000). A Rulebook for Arguments (3rd ed.). Hackett Publishing.